Susan M Rogers
May 23, 2008

Data Integrity Check for the Diabetes Prevention Project (DPP) 2008 Full Scale Release

As a partial check of the integrity of the 2008 Full Scale Release of the Diabetes Prevention Program (DPP) data archived in the NIDDK data repository, a series of tabulations was performed to verify that published results from the DPP study can be reproduced using the archived datasets. Several analyses were performed to duplicate results published by the DPP Research Group in the following publication:

> The Diabetes Prevention Program: Baseline characteristics of the randomized cohort (2000) *Diabetes Care* 23:1619-1629

The full text of the selected article can be found in Attachment 1. STATA (v10) code for our data integrity checks is included in Attachment 2.

Most of our replicated analyses produced similar results to the published tabulations. Sample sizes and tabulations of baseline sample characteristics from our analyses were similar to those in the published articles and for some subgroups the numbers matched exactly. Some discrepancies between the published and replicated analyses arose since only study clinic sites whose IRBs approved the distribution of their data to the NIDDK repository are included in the archived data, hence in some cases it is impossible to match the sample Ns; other discrepancies were related to the collapsing of some variables into wider groupings, e.g. age, BMI, race/ethnicity, when the DPP dataset was de-identified for inclusion in the NIDDK repository.

<u>The 2008 Full-Scale Data Release: Baseline and Follow-up Data</u>
The DPP Research Group reports results for 3,665 participants in the 2008 Full Scale Data Release who were randomly assigned to one of four treatment arms aimed at preventing type 2 diabetes in high-risk adults: metformin (N=1,027), troglitazone (N=584), intensive lifestyle (N=1,024), or placebo (N=1030). Eligibility criteria for the study included: age 25 years and older, a body mass index of 24 kg/m^2 or higher (22+ if Asian American), a fasting plasma glucose (FPG) level of 95-125 mg/dl (<=125 for American Indians), and a 2-hour plasma glucose concentration of 140-199 mg/dl (see 2008 Data Release Documentation for full details). Eligible participants were identified through a multi-step screening and recruitment process. Enrollment began in 1996 and participants were followed through 2001. Follow-up assessments (quarterly, semi-annual, annual, etc) included various physical measurements, medical history updates, questionnaire administration, medication adherence, and medical testing according to a standard protocol.

In de-identifying the data, all personal identifiers including all dates were removed from the DPP archive. Variables that could possibly identify a particular individual were grouped, e.g., race/ethnicity was recoded into 4 groups (Caucasian, African American, Hispanic and Other), age at baseline was recoded in 5-year groupings with truncation of those <40 and 65+, fasting glucose levels less than 100 at baseline appear as 99, and baseline BMI is provided in two

alternative groupings (2 kg/m^2 with truncation of those ≤ 26 kg/m^2 and those ≥ 42 kg/m^2 and tertiles of <30, 30 to <35, and ≥ 35 kg/m^2. Only research data are included in the released dataset (screening and post-randomization visits, clinic visits, lifestyle visits, laboratory data). Non-research data and adverse event data are not included.

Data Forms

The NIDDK data repository includes 37 data files – 31 files of data collected on standard forms ('form' files) and 6 files of data not collected on forms ('nonform' files) that include laboratory data, nutrition, quality of well-being, CT-scan, a summary of event variables for diabetes, and a baseline file (treatment assignment, baseline age and BMI group, sex, and race/ethnicity). Several variables are common to all datasets and can be used to link multiple files or to match specific visits across multiple forms. These include RELEASE_ID (unique subject ID), VISIT (baseline, screening, annual visits, interim unscheduled visit, etc), and DAYSRAND (number of days between a visit and randomization). For this verification exercise, data from 4 of the 37 files were accessed.

All form files are named with the prefix DPP_REL and either a S (screening form), F (follow-up form), TR (form for participants randomized to troglitazone), Q (questionnaires), L (lifestyle form), E (events form, e.g. pregnancy), or R (report, e.g. CHD risk status) suffix. One dataset exists for each DPP form and corresponding SAS dataset. Variables and measurement times for all non-form files, e.g., LAB, Nutrient Data, CT Scan Data, Quality of Well Being Data, Baseline Data, and Events Data, are provided in the Data Release Documentation (pages 16-26).

The Diabetes Prevention Program Research Group (2000) Baseline characteristics of the randomized cohort. *Diabetes Care* 23: 1619-1629.

The purpose of this manuscript (see Attachment 1) is to describe the baseline demographic and clinical characteristics of participants randomized to one of three treatment groups by gender, treatment assignment, and race/ethnicity. (Early results led to the discontinuation of the troglitazone treatment arm in 1998.) Of the 3,234 DPP participants randomized to either the lifestyle, metformin, or placebo treatment groups that are included in this manuscript, the NIDDK repository includes data on 3,081. Therefore, comparison of results from analyses using the archived data cannot provide a precise match to the published results. A copy of the manuscript is included in Attachment 1. STATA 10 programming code to replicate Tables 1, 3, and 4 in the published manuscript is shown in Attachment 2.

A set of initial tabulations was performed to replicate the distribution of participant characteristics across the experimental conditions as shown in Table 1 of the manuscript (**TABLE A)**. The published table describes the DPP sample at baseline by: treatment group assignment, age, gender, race/ethnicity, fasting glucose, BMI, and blood pressure. The variables from the NIDDK repository datasets that were used to replicate Table 1 are shown below. As shown in TABLE A, there is reasonable agreement between the values displayed in Table 1 of the published manuscript (left column) and values calculated from the data repository (right column) given the difference in sample sizes and the calculation of means and standard deviations from grouped variables. During de-identification of the data archive, age was recoded into the intervals: <40, 40-44, 45-49, 50-54, 55-59, 60-64, >64. To estimate the mean age (±

SD) in the archive dataset, we substituted the midpoint values for each of the age ranges and the values 35 and 70 for the endpoints. Similarly, BMI values were coded into categories to protect participant identity and we substituted the midpoint for each 2kg/m^2 grouping and 24 and 44 for the endpoints. During de-identification of the repository data, the racial/ethnic categories of American Indian and Asian-American were combined and renamed 'Other'.

NIDDK repository variables used to replicate Table 1.

Table 1 Variable	**NIDDK variable used in replication**
Treatment group assignment	DPP_REL.Basedata, assign
Age	DPP_REL.Basedata, agegroup
Sex	DPP_REL.Basedata, sex
Race/ethnicity	DPP_REL.Basedata, race_eth
Fasting glucose	DPP_REL.LAB, G000 [and visit=="BAS"]
BMI	DPP_REL.Basedata, bmi_cat
Blood pressure	DPP_REL.S03, sosbpa sodbpa

TABLE B compares results presented in Table 3 of the manuscript of self-reported characteristics of participants by sex and race/ethnicity, with calculations derived from the archived dataset. The variables from the NIDDK repository datasets that were used to replicate Table 3 are shown below. Patient characteristics include family history of type II diabetes, history of high cholesterol, history of hypertension, and among women, history of gestational diabetes and are tabulated by gender and treatment assignment. As shown in TABLE B, results obtained from the archived data closely match the published tabulations. Small discrepancies can be expected since data from 138 women and 15 men that were included in the published table are omitted from the DPP Full Scale Data Release (see 'other' racial/ethnic category). The estimates of family history of type 2 diabetes were tabulated from multiple variables found in DPP_Form S05, Part II, Participant History/Family Information. The composite variable was coded 'yes' for a family history if the respondent answered yes to the question "Did your mother or father have diabetes?" [simdiab, sifdiab] or provided a value ≥ 1 to the question "How many of your brothers and sisters have or had diabetes?" [sisibdi].

NIDDK repository variables used to replicate Table 3.

Table 3 Variable	**NIDDK variable used in replication**
Race/ethnicity	DPP_REL.Basedata, race_eth
Sex	DPP_REL.Basedata, sex
Family hx of type 2 diabetes	DPP_REL.S05, simdiab sifdia sisibdi
History of high cholesterol	DPP_REL.S05, silipi1
History of hypertension	DPP_REL.S05, sihype1
History of gestational diabetes	DPP_REL.S03, sodiab

Clinical characteristics by racial/ethnic group are presented in the published manuscript for men and women separately in Tables 4 and 5. For this exercise, we calculated characteristics for male subjects only; the variables used from the NIDDK repository to replicate Table 4 are shown below. In addition to presenting sample distributions, the published table provides means, standard deviations, and ranges for age, BMI, blood pressure, fasting plasma glucose, 2-hour

plasma glucose, HbA, fasting insulin, 30-minute insulin, fasting proinsulin, total cholesterol, HDL and LDL cholesterol, and triglycerides. Results of the calculations from the NIDDK repository are shown in **TABLE C**. Published and calculated estimates are similar, and identical for most characteristics among the Caucasian, African American and Hispanic race/ethnicity groupings. Minor variation between the published and tabulated estimates appear for insulin results, however a footnote for published Table 4 indicates that insulin results were available for only 848 of the 1,043 men. The NIDDK repository includes values for all 3 insulin measures for 1,028 men.

NIDDK repository variables used to replicate Table 4.

Table 4 Variable	**NIDDK variable used in replication**
Race/ethnicity	DPP_REL.Basedata, race_eth
Age	DPP_REL.Basedata, agegroup
BMI	DPP_REL.Basedata, bmigroup
Blood pressure	DPP_REL.S03, sosbpa sodbpa
Fasting plasma glucose	DPP_REL.LAB, G000 [visit==”BAS”]
2-hour plasma glucose	DPP_REL.LAB, G120 [visit==”BAS”]
HbA_{1c}	DPP_REL.LAB, HBA1 [visit==”BAS”]
Fasting insulin	DPP_REL.LAB, I000 [visit==”BAS”]
30-min insulin	DPP_REL.LAB, I030 [visit==”BAS”]
Fasting proinsulin	DPP_REL.LAB, PIN [visit==”BAS”]
Total cholesterol	DPP_REL.LAB, CHOL [visit==”BAS”]
HDL cholesterol	DPP_REL.LAB, CHDL [visit==”BAS”]
LDL cholesterol	DPP_REL.LAB, CLDL [visit==”BAS”]
Triglycerides	DPP_REL.LAB, TRIG [visit==”BAS”]

TABLE A. Comparison of participant characteristics by treatment group as reported in *Diabetes Care* 23(11)*:* 1620, 2000 with tabulations from the 2008 DPP Full Scale Data Release in the NIDDK repository

Table 1. Participant characteristics by treatment group assignment

	Published [a]				*Calculated from NIDDK data repository* [b]			
	Overall	**Lifestyle**	**Metformin**	**Placebo**	***Overall***	***Lifestyle***	***Metformin***	***Placebo***
n	3234	1079	1073	1082	*3081*	*1024*	*1027*	*1030*
Age (yrs)	50.6 ± 10.7	50.6 ± 11.3	50.9 ± 10.3	50.3 ± 10.4	*50.6 ± 10.4* [c]	*50.6 ± 10.8*	*50.8 ± 10.2*	*50.3 ± 10.2*
Sex								
Male	1043 (32.3)	345 (32.0)	363 (33.8)	335 (31.0)	*1028 (33.4)*	*339 (33.1)*	*358 (34.9)*	*331 (32.1)*
Female	2191 (67.7)	734 (68.0)	710 (66.2)	747 (69.0)	*2053 (66.6)*	*685 (66.9)*	*669 (65.1)*	*699 (67.9)*
Race/ethnicity								
Caucasian	1768 (54.7)	580 (53.8)	602 (56.1)	586 (54.2)	*1768 (57.5)*	*580 (56.6)*	*602 (58.6)*	*586 (56.9)*
African Amer.	645 (19.9)	204 (18.9)	221 (20.6)	220 (20.3)	*644 (20.9)*	*204 (19.9)*	*221 (21.5)*	*219 (21.3)*
Hispanic	508 (15.7)	178 (16.5)	162 (15.1)	168 (15.5)	*508 (16.5)*	*178 (17.4)*	*162 (15.8)*	*168 (16.3)*
American Ind.	171 (5.3)	60 (5.6)	52 (4.8)	59 (5.5)	*161 (5.2)* [d]	*62 (6.1)*	*42 (4.1)*	*57 (5.3)*
Asian-Amer.	142 (4.4)	57 (5.3)	36 (3.4)	49 (4.5)				
Fasting glucose (mmol/L)	5.9 ± 0.5	5.9 ± 0.4	5.9 ± 0.5	5.9 ± 0.5	*5.9 ± 0.4*	*5.9 ± 0.4*	*5.9 ± 0.4*	*5.9 ± 0.4*
BMI (kg/m^2)	34.0 ± 6.7	33.9 ± 6.8	33.9 ± 6.2	34.2 ± 6.8	*33.5 ± 5.8* [e]	*33.4 ± 5.7*	*33.4 ± 5.8*	*33.7 ± 5.9*
Blood pressure (mmHg)								
Systolic	123.7 ± 14.7	123.7 ± 14.8	124.0 ± 14.9	123.5 ± 14.4	*124.1 ± 14.7*	*124.2 ± 14.8*	*124.5 ± 14.9*	*123.8 ± 14.4*
Diastolic	78.3 ± 9.3	78.6 ± 9.2	78.3 ± 9.5	78.0 ± 9.2	*78.5 ± 9.3*	*78.8 ± 9.2*	*78.5 ± 9.5*	*78.2 ± 9.2*

Notes: Data are means ±SD or n (%) unless otherwise stated.

[a] From: The Diabetes Prevention Program: Baseline characteristics of the randomized cohort, *Diabetes Care* 23(11):1619-29, 2000

[b] Repository data include study sites whose IRBs gave permission for inclusion of participant data. Of the 3665 participants included in the 2008 Full Scale Data Release, 584 participants were randomized to the troglitazone arm of the study and not included in these tabulations. An additional 153 participants that were included in the published manuscript were excluded from the archive dataset. Tabulations derived from DPP_REL.basedata, DPP Form.S01, and DPP Form.S03.

[c] Repository data include age coded only as a categorical variable: <40, 40-44, 45-49, 50-54, 55-59, 60-64, >64. To estimate mean age ± SD, the midpoint values for each range and the values of 35 and 70 for the endpoints were substituted in the calculations.

[d] Repository data collapse American Indian and Asian Americans into a single Other category.

[e] Repository data include BMI as the categorical variable: ≤ 26, 27-28, 39-30, 31-33, 34-35, 36-37, 38-39, 40-41, ≥ 42. To estimate the mean BMI ± SD, the midpoint values for each range and the values 24 and 44 were substituted in the calculations.

TABLE B. Comparison of participant characteristics by sex and race/ethnicity as reported in *Diabetes Care* 23(11):1622, 2000 with tabulations from the 2008 DPP Full Scale Data Release in the NIDDK repository

Table 3. Self-reported characteristics of participants by sex and race/ethnicity

Original published table [a]

	All	*Caucasian*	*African Amer.*	*Hispanic*	*Amer. Indian*	*Asian Amer.*
Men						
n	1043	608	165	167	20	83
Family hx of type 2 diabetes	690 (66.2)	390 (64.3)	117 (70.9)	112 (67.1)	13 (65.0)	58 (69.9)
History of high cholesterol	389 (37.3)	234 (38.5)	65 (39.4)	53 (31.7)	3 (15.0)	34 (41.0)
History of hypertension	302 (29.0)	171 (28.1)	58 (35.2)	49 (29.3)	5 (25.0)	19 (22.0)
Women						
n	2191	1160	480	341	151	59
Family hx of type 2 diabetes	1553 (70.9)	799 (68.9)	360 (74.8)	243 (71.3)	116 (76.8)	35 (60.3)
History of gestational diabetes	353 (16.1)	191 (16.5)	63 (13.1)	55 (16.2)	36 (23.8)	8 (13.8)
History of high cholesterol	730 (33.3)	429 (37.0)	147 (30.6)	114 (33.4)	22 (14.6)	17 (29.3)
History of hypertension	569 (26.0)	303 (26.1)	144 (30.0)	68 (19.9)	40 (26.5)	15 (25.9)

Calculated from NIDDK data repository [b]

	All	*Caucasian*	*African Amer.*	*Hispanic*	*Other* [c]
Men					
n	*1028*	*608*	*165*	*167*	*88*
Family hx of type 2 diabetes	*680 (66.2)*	*390 (69.1)*	*117 (70.9)*	*112 (67.1)*	*61 (69.3)*
History of high cholesterol	*396 (38.5)*	*242 (39.8)*	*65 (39.4)*	*53 (31.7)*	*36 (40.9)*
History of hypertension	*305 (29.7)*	*175 (28.8)*	*58 (35.2)*	*50 (29.9)*	*22 (25.0)*
Women					
n	*2053*	*1160*	*479*	*341*	*73*
Family hx of type 2 diabetes	*1447 (70.5)*	*799 (68.9)*	*360 (75.2)*	*243 (71.3)*	*45 (61.6)*
History of gestational diabetes	*321 (15.6)*	*191 (16.5)*	*63 (13.2)*	*55 (16.1)*	*12 (16.4)*
History of high cholesterol	*714 (34.8)*	*434 (37.4)*	*144 (30.1)*	*115 (33.7)*	*21 (28.8)*
History of hypertension	*530 (25.8)*	*301 (26.0)*	*143 (29.9)*	*68 (19.9)*	*18 (24.7)*

[a] From: The Diabetes Prevention Program: Baseline characteristics of the randomized cohort, *Diabetes Care* 23(11):1619-29, 2000

[b] Calculations derived from 2008 Full Scale Data Release file, DPP_REL.basedata, DPP_REL.LAB, DPP_REL.S05. For comparative purposes, this analysis excludes 584 participants randomized to the troglitazone arm of the study. An additional 153 participants (15 men and 138 women) that were included in the published manuscript were excluded from the archive dataset.

[c] Repository data collapse American Indian and Asian American race/ethnic groups into a single 'Other' category.

TABLE C. Calculation of clinical characteristics in male subjects by racial/ethnic group as reported in *Diabetes Care* 23(11):1621, 2000 from the 2008 DPP Full Scale Data Release in the NIDDK repository

Calculated from NIDDK Data Repository

Table 4. Clinical characteristics in male subjects by racial/ethnic group

	All	***Caucasian***	***African Amer.***	***Hispanic***	***Other*** [a]
n	*1028*	*608*	*165*	*167*	*88*
Age at randomization (yrs) [b]					
25 to <40	*108 (10.5)*	*53 (8.7)*	*12 (7.3)*	*28 (16.8)*	*15 (17.1)*
40 to <50	*280 (27.2)*	*149 (24.5)*	*50 (30.3)*	*51 (30.5)*	*30 (34.1)*
50 to <60	*323 (31.4)*	*186 (30.6)*	*58 (35.2)*	*56 (33.5)*	*23 (26.1)*
>=60	*317 (30.8)*	*220 (36.2)*	*45 (27.3)*	*32 (19.2)*	*20 (22.7)*
BMI (kg/m^2) [b]					
<30	*447 (43.5)*	*246 (40.5)*	*66 (40.0)*	*72 (43.1)*	*63 (71.6)*
30 to <40	*497 (48.4)*	*305 (50.2)*	*84 (50.9)*	*84 (50.3)*	*24 (27.3)*
>=40	*84 (8.2)*	*57 (9.4)*	*15 (9.1)*	*11 (6.6)*	*1 (1.1)*
Blood pressure (mmHg)					
Systolic	*125.9 ± 13.9*	*125.9 ± 13.7*	*128.2 ± 14.2*	*124.0 ± 13.6*	*125.4 ± 14.3*
	80-176	*80-175*	*100-176*	*95-169*	*92-164*
Diastolic	*79.9* ± 9.3	*79.2* ± 9.6	*80.9 ± 8.9*	*80.2 ± 8.0*	*82.8 ± 9.3*
	25-105	*25-105*	*55-105*	*60-102*	*54-101*
Glycemia (mmol/L)					
Fasting plasma glucose	*6.0 ± 0.4*	*6.0 ± 0.4*	*6.0 ± 0.4*	*6.0 ± 0.5*	*6.0 ± 0.4*
	5.4-7.6	*5.4-7.6*	*5.4-7.2*	*5.4-7.6*	*5.4-7.5*
2 h plasma glucose	*9.1 ± 0.9*	*9.1 ± 0.9*	*9.0 ± 1.0*	*9.0 ± 1.0*	*9.0 ± 0.9*
	7.7-10.9	*7.7-10.9*	*7.7-10.9*	*7.7-10.9*	*7.7-10.9*
HbA_{1c} (%)	*5.9 ± 0.5*	*5.8 ± 0.4*	*6.2 ± 0.7*	*5.9 ± 0.5*	*5.9 ± 0.4*
	4-7.7	*4-7.2*	*4.2-7.7*	*4.4-7.2*	*4.8-6.8*
HbA_{1c} > 6.1%	*316 (30.7)*	*136 (22.4)*	*106 (64.2)*	*47 (28.1)*	*27 (30.7)*
Insulinemia (pmol/L)					
Fasting insulin	*159.8 ± 97.1*	*157.5 ± 99.0*	*153.8 ± 74.4*	*177.5 ± 113.6*	*153.5 ± 84.2*
	26.4-1104	*27-684*	*26.4-510*	*42.6-1104*	*36-480*
30-min insulin	*589.5 ± 402.7*	*563.2 ± 417.4*	*544.9 ± 319.9*	*689.4 ± 394.6*	*665.0 ± 419.7*
	27-4854	*31.2-4854*	*66-1812*	*27-2190*	*78-2280*
Fasting proinsulin	*20.9 ± 16.8*	*20.4 ± 17.1*	*21.5 ± 17.8*	*23.1 ±16.7*	*18.8 ± 11.8*
	2-144	*2-131*	*4.8-144*	*3.8-100*	*3.5-67*
Lipids (mmol/L)					
Total cholesterol	*5.2 ± 0.9*	*5.2 ± 0.9*	*5.2 ± 0.9*	*5.2 ± 0.9*	*5.3 ± 1.0*
	2-8.9	*2.9-8.1*	*2.0-8.9*	*2.8-7.8*	*2.9-7.6*
HDL cholesterol	*1.0 ± 0.2*	*1.0 ± 0.2*	*1.1 ± 0.2*	*1.0 ± 0.2*	*1.0 ± 0.2*
	0.5-2.2	*0.5-2.2*	*0.6-1.9*	*0.5-1.7*	*0.7-1.6*
LDL cholesterol	*3.3 ± 0.8*	*3.2 ± 0.8*	*3.4 ± 0.9*	*3.2 ± 0.9*	*3.4 ± 0.9*
	0.9-7.1	*1.0-6.0*	*0.9-7.1*	*1.0-6.3*	*1.6-5.3*
Triglycerides	*1.9 ± 1.2*	*2.0 ± 1.2*	*1.5 ± 0.8*	*2.2 ± 1.4*	*2.0 ± 1.0*
	0.4-9.4	*0.4-9.4*	*0.4-6.4*	*0.6-9.2*	*0.6-5.4*

Tabulated from 2008 DPP Full Scale Data Release, DPP_REL.Basedata, DPP_REL.S03, DPP_REL.LAB. For comparision with published data, tabulations exclude 200 male observations in data repository that were assigned to troglitazone arm of the study.

[a] Indian Americans and Asian Americans were grouped as 'Other' during de-identification of the data prior to submission to the

[b] Means, S.D., and ranges not presented. Sample *N*s for the age and BMI values (Caucasian, African American, and Hispanic) were identical to published estimates.

ATTACHMENT 1

The full text of the article referenced will be provided to approved data requestors along with the archived data.

The Diabetes Prevention Program Research Group (2000) Baseline characteristics of the randomized cohort. *Diabetes Care* 23(11): 1619-1629

ATTACHMENT 2

STATA/SE10 Code for Tabulations of Baseline Characteristics from the DPP Dataset in the NIDDK Repository
[*Diabetes Care* 23: 1619-1629, 2000; Tables 1, 3, and 4]

```
/*Tabulations from DPP repository N=3665, basedata, S01, S03
Table 1 Baseline Characteristics, Nov 2000*/

*random variable excludes Rs assigned to Troglitazone
gen random=.
replace random=1 if assign=="Lifestyle"
replace random=2 if assign=="Metformin"
replace random=3 if assign=="Placebo"
label define random 1"Lifestyle" 2"Metformin" 3"Placebo"
label val random random

***the value labels for agegroup were not explicitly defined in the repository
label define agegroup 1"<40" 2"40-44" 3"45-49" 4"50-54" 5"55-59" 6"60-64" 7"65+"
label val agegroup agegroup

gen Rage=.
replace Rage=35 if agegroup==1
replace Rage=42 if agegroup==2
replace Rage=47 if agegroup==3
replace Rage=52 if agegroup==4
replace Rage=57 if agegroup==5
replace Rage=62 if agegroup==6
replace Rage=70 if agegroup==7
label var Rage "recoded age"
tab age Rage
summarize Rage if random <=3
summarize Rage if random==1
summarize Rage if random==2
summarize Rage if random==3

label define sex 1"men" 2"women"
label val sex sex
tab sex random, col

/*value labels for race/ethnicity
American Indian subsample not included in repository*/

label define race_eth 1"Caucasian" 2"AA" 3"Hisp" 4"other"
label val race_eth race_eth
tab race_eth random, col

*baseline BMI provided as categorical variable only
gen BMIR=.
replace BMIR=24 if bmi_cat==1
replace BMIR=27 if bmi_cat==2
replace BMIR=29 if bmi_cat==3
replace BMIR=31 if bmi_cat==4
replace BMIR=33 if bmi_cat==5
replace BMIR=35 if bmi_cat==6
replace BMIR=37 if bmi_cat==7
replace BMIR=39 if bmi_cat==8
replace BMIR=41 if bmi_cat==9
replace BMIR=44 if bmi_cat==10
tab bmi_cat BMIR
summarize BMIR if random <=3
sort random
by random: summarize BMIR
```

```
*blood pressure
summarize sosbpa if random <=3
by random: summarize sosbpa
summarize sodbpa if random <=3
by random: summarize sodbpa


****Table 3. Gestational diabetes
tab sodiab
gen gestdiab=.
replace gestdiab=1 if sodiab==2 & sex==2
replace gestdiab=0 if sodiab!=2 & sex==2
tab gestdiab
tab gestdiab if random<=3
tab gestdiab race_eth if random<=3, col


**********Table 4 calculations using this dataset--need to merge lab and
basedata for other clin. characteristics
gen Ragegrp=.
replace Ragegrp=1 if agegroup==1
replace Ragegrp=2 if agegroup==2 | agegroup==3
replace Ragegrp=3 if agegroup==4 | agegroup==5
replace Ragegrp=4 if agegroup==6 | agegroup==7
label define Ragegrp 1"<40" 2"40-49" 3"50-59" 4"60+"
label val Ragegrp Ragegrp

sort race_eth
tab Ragegrp if sex==1 & random<=3
by race_eth: tab Ragegrp if sex==1 & random<=3

label define bmi_cat 1"<=26" 2"26-28" 3"28-30" 4"30-32" 5"32-34" 6"34-36" 7"36-
38" 8"38-40" 9"40-42" 10"42+"
label val bmi_cat bmi_cat
gen RBMI=.
replace RBMI=1 if bmi_cat==1 | bmi_cat==2 | bmi_cat==3
replace RBMI=2 if bmi_cat>=4 & bmi_cat<=8
replace RBMI=3 if bmi_cat==9 | bmi_cat==10
label define RBMI 1"<30" 2"30<40" 3"40+"
label val RBMI RBMI
tab bmi_cat
tab RBMI
tab RBMI if sex==1 & random<=3
by race_eth: tab RBMI if sex==1 & random<=3

summarize sosbpa if sex==1 & random<=3
summarize sodbpa if sex==1 & random<=3
by race_eth: summarize sosbpa if sex==1 & random<=3
summarize sodbpa if sex==1 & random<=3
by race_eth: summarize sodbpa if sex==1 & random<=3
summarize sodbpa if sex==1 & random<=3


**********TABLE 3
/*Tabulations from DPP repository N=3665, basedata S05
several observations removed by DCC during de-identification of data
```

```
Table 3 Baseline Characteristics, Nov 2000*/

*random variable excludes Rs assigned to Troglitazone
gen random=.
replace random=1 if assign=="Lifestyle"
replace random=2 if assign=="Metformin"
replace random=3 if assign=="Placebo"
label define random 1"Lifestyle" 2"Metformin" 3"Placebo"
label val random random

label define sex 1"men" 2"women"
label val sex sex
tab random sex, row

label define race_eth 1"Caucasian" 2"AA" 3"Hisp" 4"other"
label val race_eth race_eth
tab sex race_eth if random <=3

*FORM S05 hx of high blood pressure, high cholesterol
label define yesno 1"yes" 2"no"
label val sihype1 yesno
label val silipi1 yesno
label val sihype2 yesno
label val silipi2 yesno

tab sihype1 sihype2
tab silipi1 silipi2

sort sex
by sex: tab sihype1 race_eth if random <=3, col
by sex: tab silipi1 race_eth if random <=3, col


*create variable family history of diabetes Form S05
tab simdiab
tab sifdia
tab sisibdi
gen famdiab=.
replace famdiab=1 if simdiab==1 | sifdia==1 | sisibdi >=1
replace famdiab=0 if simdiab !=1 & sifdia !=1 & sisibdi==0
tab simdiab sifdia
label val famdiab yesno
tab famdiab
by sex: tab famdiab race_eth if random <=3, col


*******TABLE 4
/*Tabulations from DPP repository N=3665, merge basedata and lab data use many
to one option
select visit="BAS" & sex==1, use variable 'random' to exclude Trog arm for this
analysis
Table 4 Male Baseline Characteristics, Nov 2000*/

*random variable excludes Rs assigned to Troglitazone
gen random=.
replace random=1 if assign=="Lifestyle"
replace random=2 if assign=="Metformin"
```

```
replace random=3 if assign=="Placebo"
label define random 1"Lifestyle" 2"Metformin" 3"Placebo"
label val random random

label define race_eth 1"Caucasian" 2"AA" 3"Hisp" 4"other"
label val race_eth race_eth
tab race_eth if visit=="BAS" & sex==1 & random <=3

sort race_eth

*To convert mg/dl of glucose to mmol/l, divide by 18 or multiply by 0.055.
gen G000R=g000*.055
summarize G000R if visit=="BAS" & sex==1 & random <=3
by race_eth: summarize G000R if visit=="BAS" & sex==1 & random <=3

gen G120R=g120*.055
summarize G120R if visit=="BAS" & sex==1 & random <=3
by race_eth: summarize G120R if visit=="BAS" & sex==1 & random <=3

summarize hba1 if visit=="BAS" & sex==1 & random <=3
by race_eth:summarize hba1 if visit=="BAS" & sex==1 & random <=3

gen hba6=.
replace hba6=1 if hba1>=6.1
replace hba6=0 if hba1<6.1
tab hba6 if visit=="BAS" & sex==1 & random <=3
by race_eth: tab hba6 if visit=="BAS" & sex==1 & random <=3

*insulin
*glucose conversion uU/mL to pmol/L multiply by 6
*gen i000R=i000*6
summarize i000R if visit=="BAS" & sex==1 & random <=3
by race_eth:summarize i000R if visit=="BAS" & sex==1 & random <=3

*gen i030R=i030*6
summarize i030R if visit=="BAS" & sex==1 & random <=3
by race_eth:summarize i030R if visit=="BAS" & sex==1 & random <=3

summarize pin if visit=="BAS" & sex==1 & random <=3
by race_eth:summarize pin if visit=="BAS" & sex==1 & random <=3

*lipids
/*To convert mmol/l of HDL or LDL cholesterol to mg/dl, multiply by 39
To convert mg/dl of HDL or LDL cholesterol to mmol/l, divide by 39

To convert mmol/l of triglycerides to mg/dl, multiply by 89
To convert mg/dl of triglycerides to mmol/l, divide by 89

To convert umol (micromoles) /l of creatinine to mg/dl, divide by 88
To convert mg/dl of creatinine to umol/l, multiply by 88*/

gen cholrec=chol/38.67
*total cholesterol mg dL to mmol L divide by 38.67
summarize cholrec if visit=="BAS" & sex==1 & random <=3
by race_eth:summarize cholrec if visit=="BAS" & sex==1 & random <=3

*HDL
```

```
*gen chdlR=chdl/38.67
summarize chdlR if visit=="BAS" & sex==1 & random <=3
by race_eth:summarize chdlR if visit=="BAS" & sex==1 & random <=3

*LDL
*gen cldlR=cldl/38.67
summarize cldlR if visit=="BAS" & sex==1 & random <=3
by race_eth:summarize cldlR if visit=="BAS" & sex==1 & random <=3

*gen trigR=trig/89
summarize trigR if visit=="BAS" & sex==1 & random <=3
by race_eth:summarize trigR if visit=="BAS" & sex==1 & random <=3

****For Table 1 glucose
sort random
by random:summarize G000R if visit=="BAS"
```